Large scale visual odometry using stereo vision

Andres Hernandez-Gutierrez[1], Juan I. Nieto[1], Teresa Vidal-Calleja[1,2], Eduardo Nebot[1]
[1]Australian Centre for Field Robotics
The University of Sydney, NSW 2006 ,Australia

[2]Institut de Robòtica i Informàtica Industrial
CSIC-UPC, Llorens Artigas 4-6 - 08028, Barcelona, Spain
{a.hernandez, j.nieto, t.vidal, e.nebot}@acfr.usyd.edu.au

Abstract

This paper presents a system for egomotion estimation using a stereo head camera. The camera motion estimation is based on features tracked along a video sequence. The system also estimates the tridimensional geometry of the environment by fusing the visual information from multiple views. Furthermore, the paper presents comparisons between two different algorithms. The first one is by applying triangulation to 3D points. Motion estimation using 3D points suffers from the problem of non-isotropic noise due to the large uncertainty in depth estimation. To deal with this problem we present results with a second approach that works directly in the disparity space. Experimental results using a mobile platform are presented. The experiments cover long distances in urban-like environments with the presence of dynamic objects. The system presented is part of a bigger project involving autonomous navigation using vision only.

1 Introduction

A common aspect in most perception or control mobile platform architectures is the need for a reliable motion estimation. Egomotion estimation has been tackled using various approaches and different sensing modalities. Common approaches found in literature are based on using GPS, inertial measurement units (IMU), encoders, or systems based on the fusion of these sensors. Although GPS sensors have become very popular due to the fact that they can provide accurate localisation, they are not appropriate for reliable localisation under non-free sky conditions.

Vision systems play an important role in autonomous navigation. The richness of the information acquired, relative low cost and weight of cameras make these sensors highly attractive. Despite these positive remarks, there are several factors that make vision a hard problem. Firstly of all, the complexity of the real world is significantly superior to the complexity of the information acquired by a camera. Secondly, the pixel value recorded by a camera depends not only on the shape of the observed object, but also on the illumination and dynamics of the environment [Ma *et al.*, 2006]. Even though there are still many unsolved problems, such as dealing with different light conditions. In contrast, vision has been widely adopted in navigation systems for autonomous vehicles to perform different tasks. Examples include object avoidance, motion detection and classification [Agrawal *et al.*, 2005], localisation [Nister *et al.* , 2004] and [Agrawal and Konolige, 2006] where the latter authors integrate information from a GPS sensor and wheel encoders to perform the egomotion estimation. Besides, 3D environment reconstruction [Mouragnon *et al.*, 2006] and other important aspects related to creating autonomous vehicles capable of navigating not only in unstructured scenes, but also in dynamic environments.

We present here results for motion estimation using stereo vision only. The system presented is part of a bigger project which aims for autonomous navigation using vision sensors only. The system is based on feature detection and tracking between consecutive frames. Features are extracted from both the left and right images of the stereo head and matched in a pairwise manner. Due to the presence of outliers in the matches, a robust estimation process is needed. We use RANSAC in order to select the inliers from the putative correspondences. This stage is followed by the triangulation process which computes the 3D point location represented by the feature in the pair of stereo images. Subsequently, features from the previous left image are tracked in the left image of the next frame to estimate the rotation matrix and translation vector between the two consecutive frames. This approach is based on the 3D point locations estimated by the stereo head.

Unfortunately, visual odometry based on 3D points suffers from the problem of non-isotropic noise due to

the large uncertainty in depth estimation. For this reason, we also provide results of an approach working in the disparity space. The algorithm calculates a homography matrix with disparity images. The homography matrix is computed using the rotation matrix and the translation vector based on the 3D point correspondences from two consecutive left images. Following this step, a non-linear minimisation process using the Levenberg-Marquardt algorithm is used to minimise the distance between the reprojected features and the measured ones.

The paper is organised as follows. Firstly, the feature extraction and matching processes are explained in section 2. Details about how to recover the rotation and translation matrix are given in section 3, where the theory about estimates using the disparity space are also presented. This section is followed by the optimisation process in section 4. Finally, experimental results and conclusions are given in sections 5 and 6 respectively.

2 Feature extraction and feature matching

The motion estimation process is based on a 3D model that represents the extracted features from the stereo images. There exist different methods to obtain these features, such as Harris corner detector [Harris and Stephens, 1988], which has been the most common approach to perform this task. However, features extracted using this detector are not scale or viewpoint invariant. In order to overcome this issue, we used instead SIFT key-points [Lowe, 2004] that were designed to be invariant to rotation, scaling and small changes in viewpoint.

Once the features from the left and right images are extracted, they are matched to each other using the nearest-neighbor algorithm. Because after this stage some mismatches could emerge, an appropriate approach to reject outliers is needed. We use RANSAC algorithm that takes into account outliers along with the fundamental matrix. The robust matching process is accomplished as follows:

1. A sample of 8 points is selected randomly to compute the fundamental matrix.
2. This fundamental matrix is applied to all features in the left image, so that they are reprojected into the right image.
3. The features are considered as inliers for that fundamental matrix obtained, if the distance from the reprojected to the measured ones in the right image is less than a threshold value previously defined.
4. Steps from 1 to 3 are repeated a fixed number of times and the fundamental matrix hypothesis having the greater number of inliers is kept.

Once the features are matched, the disparity (which will allow us to compute the 3D model) has to be calculated. Only features having enough disparity (*e.g.* values greater than 5 pixels) are considered, as the depth error increases with respect to the range to the 3D feature location.

3 Motion estimation

In order to estimate the 3D location of the features, the remainder matches from the left and right image (after applying the disparity filter) are triangulated using the following equations,

$$X = \frac{\hat{x}Z}{f} \tag{1}$$

$$Y = \frac{\hat{y}Z}{f} \tag{2}$$

$$Z = \frac{Bf}{d} \tag{3}$$

where $(\hat{x}, \hat{y})$ is the pixel coordinate with respect to the image centre, f the focal length, and B the baseline of the camera. The 3D feature location represented by the pixel in the image is then given by $[X, Y, Z]^T$.

For two consecutive left images, we apply the same matching procedure as in Section 2 for the stereo pair. Let $L_k = [X_k, Y_k, Z_k]^T$ and $L_k^{'} = [X_k^{'}, Y_k^{'}, Z_k^{'}]^T$ be the 3D feature locations representing the features from the previous and the current frame respectively of the k-th feature. Then, we are interested in estimating a rotation matrix R and a translation vector t such that $L^{'} = RL + t$.

3.1 R and t recovery

As in [Umeyama, 1991], the rotation matrix and the translation vector can be computed as follows:

- The centroid for each data set is calculated:

$$\mu_L = \frac{1}{N}\sum_{k=1}^{N} L_k \tag{4}$$

$$\mu_{L'} = \frac{1}{N}\sum_{k=1}^{N} L_k^{'} \tag{5}$$

- The variance is also obtained as follows:

$$\sigma_L = \frac{1}{N}\sum_{k=1}^{N} |||L_k - \mu_L||| \tag{6}$$

$$\sigma_{L'} = \frac{1}{N}\sum_{k=1}^{N} |||L_k^{'} - \mu_{L'}||| \tag{7}$$

- Then, we compute the next operation as:

$$\Sigma_{LL'} = \frac{1}{N}(L_k^{'} - \mu_{L'})(L_k - \mu_L)^T \tag{8}$$

Finally, we obtain the SVD decomposition of $\Sigma_{LL'}$ and recover the rotation matrix and translation vector from the following equations:

$$R = USV^T \tag{9}$$

$$t = \mu_{L'} - R\mu_L\,, \tag{10}$$

R and t being the estimated rotation matrix and the translation vector between two consecutive frames based on the 3D point location.

3.2 Disparity space

When computing R and t based on the 3D point location, the estimated values are affected by the uncertainty introduced in the triangulation process. In order to alleviate this issue, we work on the disparity space and estimate the homography matrix that relates image coordinates and disparity from the previous to the current frame.

First, let us define a transformation from the 3D point location $\mathbf{W} = [X, Y, Z]$. $\mathbf{W}$ to $\mathbf{W}^{'} = [X^{'}, Y^{'}, Z^{'}]$ as:

$$\begin{pmatrix} \mathbf{W}^{'} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{W} \\ 1 \end{pmatrix} \tag{11}$$

Where $\mathbf{W}$ is any extracted feature from the previous frame and $\mathbf{W'}$ its correspondence in the current frame. From Eq.(1) to Eq.(3) and including the disparity value, we obtain:

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ d \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 0 & fB \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{12}$$

Where $\mathbf{x} = [\hat{x}, \hat{y}, d]^T$ are the features coordinates in the previous left image. Then $\mathbf{M}$ and $\mathbf{W}$ are defined such that:

$$\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{W} \\ 1 \end{pmatrix} \tag{13}$$

$\mathbf{M}$ being the right side matrix multiplying the 3D feature location $[X, Y, Z]^T$. This matrix M contains the intrinsic parameters of the camera . We use a similar notation to express the feature coordinates in the current left image as:

$$\begin{pmatrix} \mathbf{x}^{'} \\ 1 \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{W}^{'} \\ 1 \end{pmatrix} \tag{14}$$

Substituting Eq.(11) in Eq.(14) we have:

$$\begin{pmatrix} \mathbf{x}^{'} \\ 1 \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{W} \\ 1 \end{pmatrix} \tag{15}$$

Now, inverting Eq.(13) and substituting in Eq.(15) we obtain the relationship between image coordinates from the previous frame and those from the current frame as follows:

$$\begin{pmatrix} \mathbf{x}^{'} \\ 1 \end{pmatrix} = \mathbf{M} \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \mathbf{M}^{-1} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \tag{16}$$

Finally, the homography matrix will not only depend on the rotation matrix and the translation vector, but also on the camera parameters that are given in matrix M. Thus, we can express this homography matrix as:

$$\mathbf{H(R,t)} = \mathbf{M} \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \mathbf{M}^{-1} \tag{17}$$

and Eq.(16) is expressed in terms of this homography matrix as follows:

$$\begin{pmatrix} \mathbf{x}^{'} \\ 1 \end{pmatrix} = \mathbf{H(R,t)} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \tag{18}$$

Therefore, we can establish a relationship in the disparity space, from features extracted in both consecutive frames, through the *disparity space homography matrix*.

4 Refinement process

In this work, we compare the motion estimation results in three different manners. Firstly, we obtain a rotation matrix and a translation vector based on the whole set of 3D points matched between two consecutive frames. Secondly, we refine the disparity space homography obtained with the rotation and translation computed as in the previous case (with the whole set of 3D points). Finally, in the third option we refine the disparity space homography obtained selecting randomly 7 points only from the data set. In the two latter cases, the estimated homography matrix is applied to all the features in the previous left image to get the reprojection on the current left image. Therefore, the reprojected feature $\mathbf{x}^{''}$ is represented as,

$$\begin{pmatrix} \mathbf{x}^{''} \\ 1 \end{pmatrix} = \mathbf{H(R,t)} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \tag{19}$$

The refinement process, required for two of the motion estimation procedures, is done using a nonlinear minimisation algorithm, such as the Levenberg-Marquardt method. The objective is to minimise the distance error between the reprojected features $\mathbf{x}^{''}$ and the measured features $\mathbf{x}^{'}$ in the current frame. Consequently, the cost function to minimise is given by:

$$\min \sum_{k=1}^{N} d(\mathbf{x}^{'}, \mathbf{x}^{''}) \tag{20}$$

After the nonlinear minimisation process, we want to recover the translation and rotation from the resulting disparity space homography matrix. Instead of working with the full homography matrix because the rotation matrix can minimally represented by 3 angles, we express the rotation in Euler angles. Thus, the Levenberg-Marquardt algorithm will provide the translation and Euler angles vector that minimise the aforementioned cost function. In contrast, this nonlinear minimisation procedure converges after 40 iterations which take about 45 ms . However, this number of iterations will depend on the number of matches as well as on the features location in the image.

5 Experimental results

The results provided in this section show the egomotion estimati on for a vehicle that was driven at about 30 km/h. The stereo camera used in this practical implementation has a baseline of 24 cm. It was placed on top in the middle of the vehicle as it can bee seen in Figure 1.[1]. It captured 640x480 pixels stereo images at a frame rate of $3Hz$.

Figure 1: Vehicle used to perform the localisation

The navigated distance was about 1.0 km in an urban-like environment. Therefore, the algorithm also has to deal with dynamic objects. Figure 2 depicts the 2D localisation of the vehicle for the trajectory obtained using the complete data set of 3D points matched between two consecutive frames.

As it can be observed in Figure 3 at frame 48, there was a significant change in the relative orientation that affected the global localisation. However, after the optimisation process, the egomotion estimation was improved. Furthermore, Figure 3 also shows small changes in orientation along the trajectory. Although these fluctuations seem to be insignificant, they affect the global rotation angle. The reason is because the orientation error is accumulated in time as it can be observed in Figure 5.

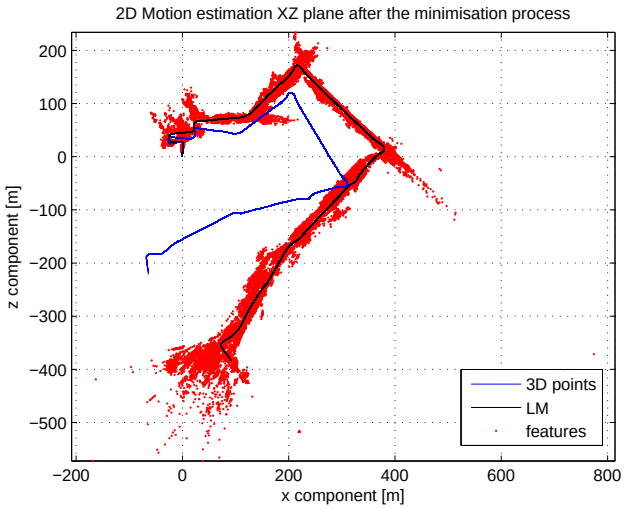

Figure 2: 2D localisation; blue line: localisation based on the 3D points cloud only; black line: localisation after the optimisation algorithm; red points: features observed along the trajectory

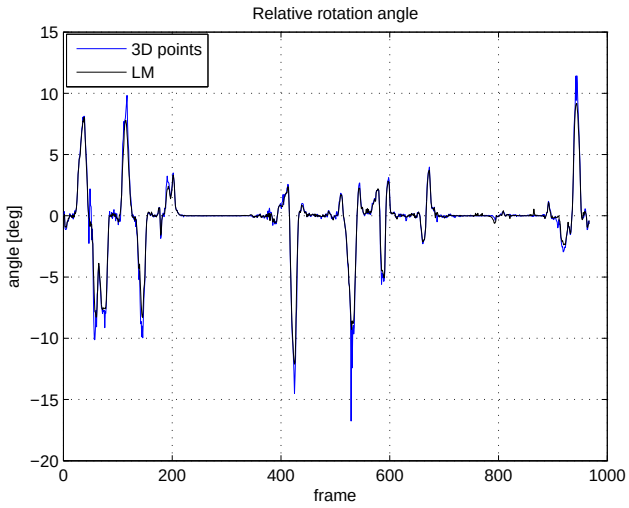

Figure 3: Relative orientation angle between two consecutive frames

These orientation errors are influenced by different factors, such as the number and distribution of matches found in the previous and the current left image as well

[1] See available video at http://www-personal.acfr.usyd.edu.au/ahernandez/egomotion_estimation.avi

Figure 4: Inlying matches between two consecutive frames used to compute the R and t parameters

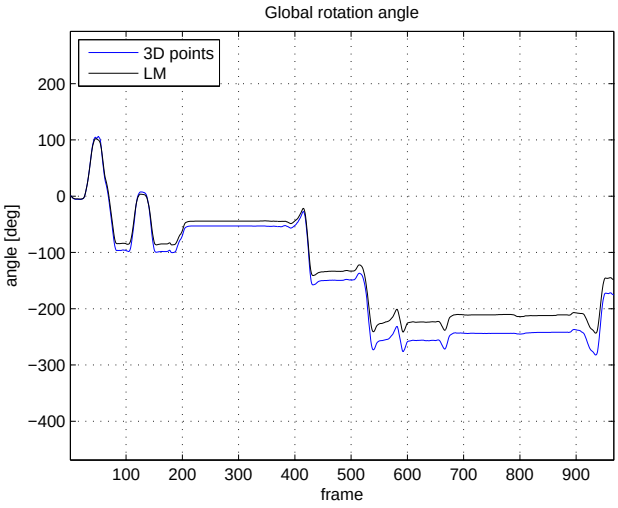

Figure 5: Global orientation angle

as the distance at which the 3D point representation of these features are located with respect to the camera position; this effect is depicted in Figure 4. On the other hand, 3D points close to the camera will enable the algorithm to estimate an appropriate translation estimation; whereas, features far from the stereo camera will allow to obtain better rotation estimation.

Besides, another important aspect that affects the vehicle localisation is the number of turns achieved along the trajectory since the orientation errors affect the vehicle localisation in the rest of the trajectory.

In order to compare the motion estimation using both methods (using the 3D points locations only and applying the optimisation method to refine the rotation and the translation vector between two frames), the generated trajectories are superimposed on a satellite view as it is detailed in Figure 6. In this case, the path with label *3D points* represents the vehicle localisation based on the 3D points locations; whereas the trajectory given by the label *LM* was obtained after applying the optimisation process.

From Figure 6, it is clear that the localisation before taking the first roundabout is better estimated after optimising the R and t parameters. However, along the trajectory estimated on the bottom of the satellite view, the localisation of the vehicle was not much improved by refining the rotation and translation parameters. Nevertheless, in the last part of the path, the motion estimation is improved using the optimisation process. Note that the final position estimated is much closer than the one using the set of 3D points only.

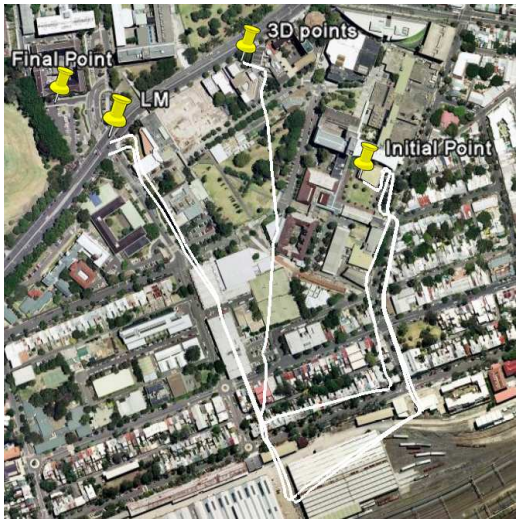

Figure 6: Satellite view using Google map to compare the vehicle localisation along the driven trajectory

The third experiment was conducted selecting only a random sample of 7 points from the whole set of 3D points. This approach allowed us to decrease the computational cost necessary to estimate the rotation and translation. Selecting only these points from the complete data set provides correct results as it can be seen in Figure 6 by the line with no label. Because the 7 points are sampled randomly, the first parameters R and t estimated could lead to a significant large reprojection error. It could also lead to a bad initialisation for the disparity space homography matrix computed using these values and applied to those features from the previous left im-

age. As a result, the number of iterations required by the optimisation process to minimise this error would increase in order to provide the best rotation matrix and translation vector that describe the motion between those two frames. Therefore, there exists two important factors to be considered. Firstly, a trade-off between accuracy and computational cost when the localisation is obtained based on the complete data set or based on a set of seven points sampled randomly. Secondly, an appropriate selection of the seven points sampled from the complete data set that will lead to a decrease in the number of iterations required by the Levenberg-Marquardt algorithm.

6 Conclusions

In this paper a system for egomotion estimation based on features extracted from a video sequence was presented. First, the motion estimation was obtained using the 3D point location represented by features tracked in a video sequence. In this case, the vehicle localisation was seriously affected by the depth error when the triangulation process was accomplished. In order to alleviate this issue, a second method, which was based on the disparity space and optimisation algorithm such as the Levenberg-Marquardt method, was presented. This led to better results in the localisation of the vehicle which was conducted over a long distance in an urban-like environment.

We proved that the egomotion estimation can be computed by selecting randomly a sample of seven points from the complete data set of 3D points. Although it provided appropriate outcomes while the computational cost was reduced, a more suitable method to extract these samples and to reduce the number of iteration in the optimisation process should be adopted. Finally, these three results were compared by displaying the generated trajectory on a satellite view of the area where the vehicle was driven.

7 Future Work

As a future work, different environment representations should be studied in order to mitigate the errors in the egomotion estimation. As a result, these representations would allow us to close loops and to decide whether the vehicle has previously visited a certain place. Moreover, this localisation method based on stereo vision will incorporate information from other sensors that will make the autonomous vehicle capable of navigate in different environments. Finally, colour information provided by the stereo camera will be used to model the terrain that will be navigated by the mobile platform.

8 Acknowledgements

This work has been supported in part by CONACYT and SEP Mexico, the Rio Tinto Centre for Mine Automation, the ARC Centre of Excellence programme, funded by the Australian Research Council (ARC) and the New South Wales State Government, and the Spanish Ministry of Innovation and Science.

References

[Ma *et al.*, 2006] Yi Ma, Stefano Soatto, Jana Kosecka and S. Shankar Sastry. An Invitation to 3-D Vision. From Images to Geometric Models. Springer, 2006.

[Agrawal *et al.*, 2005] Motilal Agrawal, Kurt Konolige, and Luca Iocchi. Real-Time detection of independent motion using stereo. In *IEEE workshop on Motion WACV/MOTION, 2005.*

[Nogueira *et at.*, 2008] Sergio Nogueira, Yassine Ruichek, and Francois Charpillet. A Self Navigation Technique using Stereovision Analysis. In *Stereo Vision Book*, Vienna, Austria, November 2008.

[Harris and Stephens, 1988] C. Harris and M. Stephens. A combined corner detector and edge detector. In *Alvey Vision Conference*, pages 147-151, 1988.

[Nister *et al.* , 2004] David Nister, Oleg Naroditsky, and James Bergen. Visual Odometry. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, pages 652–659, Los Alamitos, CA, USA, 2004.

[Valgren and Lilienthal, 2007] Christoffer Valgren and Achim Lilienthal. SIFT, SURF and Seasons: Long-term Outdoor Localization Using Local Features. In *Proc. of 3rd European Conference on Mobile Robots.* Freiburg, Germany, 2007.

[Derpanis and Chang, 2006] Konstantinos G. Derpenis and Peng Chang. Closed-form Linear Solution To Motion Estimation In Disparity Space. In *Intelligent Vehicles Symposium.* Tokyo, Japan, June, 2006.

[Agrawal and Konolige, 2006] Motilal Agrawal and Kurt Konolige. Real-Time Localization in Outdoor Environment using Stereo Vision and Inexpensive GPS. In *18th International Conference on Pattern Recognition ICPR 2006*, (3):1063–1068, 2006.

[Demirdjian and Darrell, 2001] D. Demirdjian and T. Darrell. Motion Estimation from Disparity Images. In *Proceedings of the Eighth IEEE International Conference in Computer Vision*, (1):213–218, Vancouver, BC, Canada 2001.

[Umeyama, 1991] Shinji Umeyama. Least-Squares Estimation of Transformation Parameters Between Two Patterns. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 4(13):376–380, April, 1991.

[Fischler and Bolles, 1981] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Application s to Image Analysis and Automated Cartography. *Commun. ACM.*, 6(24):381–395, June, 1981.

[Lowe, 2004] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, (60):91–110, 2004.

[Mouragnon *et al.*, 2006] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser and P. Sayd. Real Time Localization and 3D Reconstruction. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* (1):363–370. 2006.

[Point Grey Research Inc., 2004] Point Grey Research, Inc. Stereo Accuracy and Error Modeling. April. 2004.